# Содержание

# 1 Введение

Данная лабораторная работа предназначена для изучения основных информационных характеристик дискретных источников. В качестве исходного материала в работе используются тексты на естественных языках: русском, английском, французском и немецком. Программа позволяет наблюдать гистограммы одномерных распределений, рассчитывать энтропию и условную энтропию текста при различных значениях глубины информационных связей и генерировать на основе этих расчётов случайные тексты. Отдельным пунктом в программу включен опыт Шеннона для определения энтропии языка.

# 2 Интерфейс пользователя

На рис. 1 показано основное окно программы после её запуска.

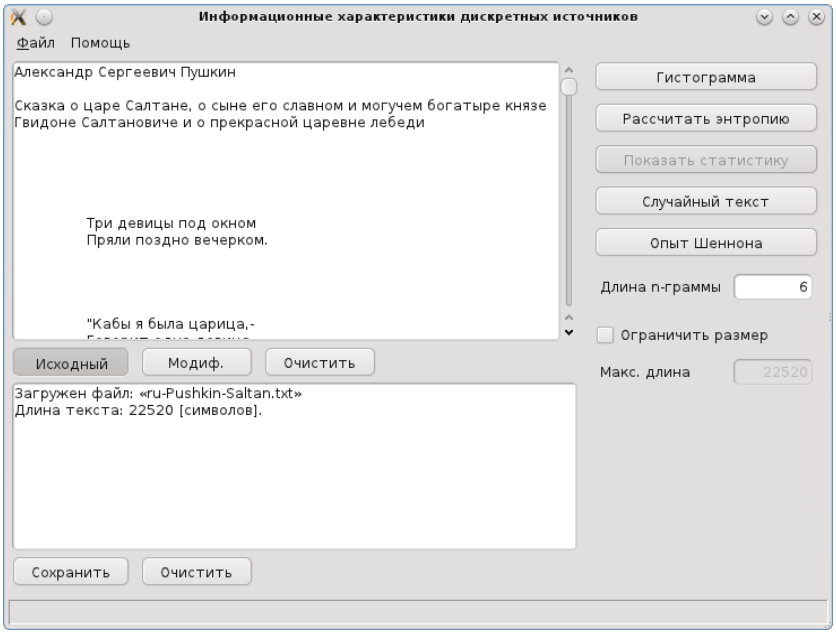



Рис. 1. Окно программы

Элементы управления включают главное меню (см. п. 2.1) и кнопки действий:

- **Гистограмма** — рассчитывает и отображает гистограмму одномерного распределения вероятностей букв в тексте (см. п. 2.2).
- **Рассчитать энтропию** — рассчитывает энтропию и выводит результат (и сопутствующие данные) в окно журнала (см. п. 2.3).
- **Показать статистику** — показывает статистику распределения $n$-грамм (см. п. 2.4).
- **Случайный текст** — генерирует некоторое количество символов случайного текста и отображает их в окне исходного текста (см. см. п. 2.5).

- **Опыт Шеннона** — открывает диалоговое окно опыта Шеннона (см. п. 2.6).

Кнопка «Исходный» под окном исходного текста отображает исходный текст. Кнопка «Модиф.» отображает модифицированный текст: текст, состоящий только из строчных букв и пробелов (акценты удаляются, 'ъ' заменяется на 'ь').

Кнопки «Очистить» очищают окна исходного текста и журнала соответственно. Кнопка «Сохранить» под окном журнала позволяет сохранить результаты расчётов из журнала в текстовый файл.

Параметр «Длина n-граммы» определяет глубину информационных связей для расчёта условной энтропии, генерации случайных текстов и опыта Шеннона. При $n = 1$ рассчитывается энтропия побуквенного распределения, связи между символами не учитываются вообще, при $n = 2$ — энтропия распределения биграмм, т.е. учитывается один предыдущий символ и т.д.

Параметр «Ограничить размер» позволяет включить ограничение размера анализируемого текста. Размер текста (в символах) задаётся в поле «Макс. длина».

## 2.1 Главное меню

- **Файл**
  - **Открыть** (`Ctrl+L`): загрузить исходный текст из текстового (txt) файла. Unicode–версия программы работает только с текстами в кодировке UTF-8.
  - **Выход** (`Alt+F4`): выйти из программы.
- **Помощь**
  - **Руководство пользователя** (`F1`): открывает это руководство.
  - **О программе**: показывает информацию о программе.

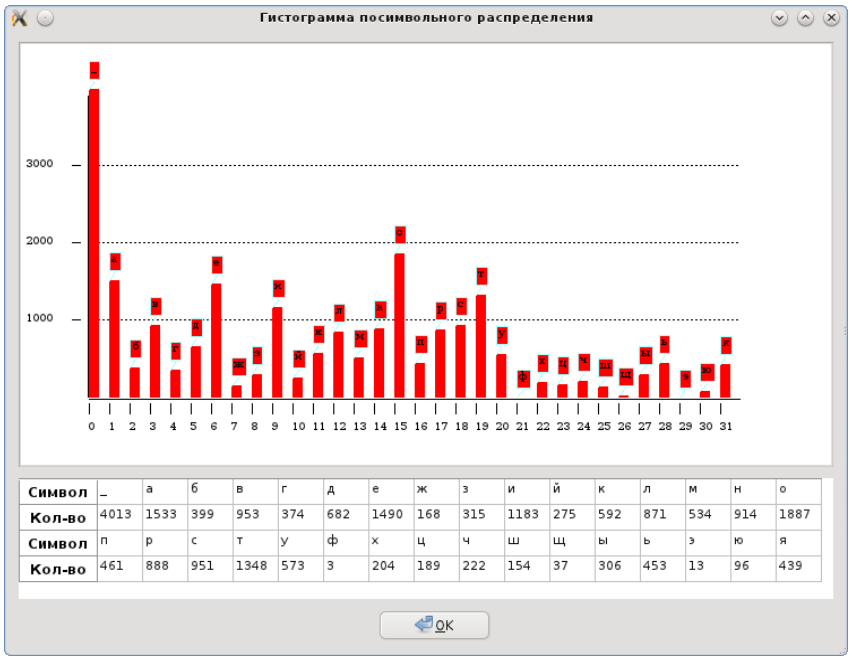
## 2.2 Гистограмма распределения

Типичный вид окна гистограммы распределения показан на рис. 2.

По оси абсцисс отложены порядковые номера символов алфавита от 0 до 31 (для русского языка): 31 буква, исключая «ё» и «ъ», плюс пробел (отображается символом «_»). По оси ординат отложено количество встречаемости каждого из символов в исходном тексте. Под гистограммой показана таблица распределения встречаемости для каждого символа.

## 2.3 Расчёт энтропии

Расчёт энтропии производится с учётом заданной длины $n$-граммы (подстроки текста длиной $n$). Типичный результат расчёта энтропии:




| Символ | _ | а | б | в | г | д | е | ж | з | и | й | к | л | м | н | о |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Кол-во | 4013 | 1533 | 399 | 953 | 374 | 682 | 1490 | 168 | 315 | 1183 | 275 | 592 | 871 | 534 | 914 | 1887 |
| Символ | п | р | с | т | у | ф | х | ц | ч | ш | щ | ы | ь | э | ю | я |
| Кол-во | 461 | 888 | 951 | 1348 | 573 | 3 | 204 | 189 | 222 | 154 | 37 | 306 | 453 | 13 | 96 | 439 |

Рис. 2. Гистограмма распределения

```
Входной файл: «ru-Pushkin-Saltan.txt»
Длина n-граммы: 6 [символов]
Общее число n-грамм: 22515
Число уникальных n-грамм: 12395
Энтропия n-граммы: H(An) = 13.11 [бит/n-грамму]
Удельная энтропия: H(An)/n = 2.19 [бит/символ]
```

Зная энтропию $n$-граммы и $(n-1)$-граммы можно определить условную энтропию:

$$H(A|A'_{n-1}) = H(A_n) - H(A_{n-1}).$$

Из теории известно, что при увеличении $n$ удельная энтропия стремится к условной и характеризует **энтропию языка** [1, 2].

## 2.4 Статистика распределения $n$-грамм

Кнопка «Показать статистику» становится активной после расчёта энтропии. Нажатие кнопки выводит таблицу распределения, где показаны все найденные $n$-граммы и их количество.

## 2.5 Случайный текст

Случайный текст генерируется с учётом заданной глубины $d$ на основе исходного текста.

Идея получения случайного текста с заданной глубиной информационных связей взята из [1]. В исходном тексте ищется отрезок длиной $n-1$. На выход выдаётся следующий за ним символ. После этого первый символ отрезка отбрасывается,

а в конец добавляется только что выданный символ. Новый отрезок затем ищется в оставшейся части текста. И так далее.

При динамическом изменении $n$ программа старается, если это возможно, учесть уже сгенерированные символы.

## 2.6 Опыт Шеннона

Оригинальный способ определения энтропии языка, предложенный в 1951 г. Шенноном, заключается в отгадывании $n$-й буквы текста при известных $n-1$ предыдущих. Мера степени неопределённости данного опыта является оценкой сверху условной энтропии [2].

В работе выбирается случайный отрезок исходного текста длины $n-1$ и пользователю предлагается угадать следующую букву. Многократное повторение опыта даёт распределение частот правильного угадывания: частоты (вероятности) $w_1, w_2, \ldots, w_K$ того, что буква будет правильно угадана с $1, 2, \ldots, K$-й попытки ($K$ — объём алфавита). Эти вероятности являются оценкой вероятностей символов алфавита, расположенных в порядке убывания частот [2]. Отсюда следует, что энтропия данного распределения будет являться оценкой (сверху) условной энтропии

$$H(A|A'_{n-1}) \leqslant H(W) = -\sum_{i=1}^{K} w_i \log w_i.$$

Диалоговое окно опыта Шеннона показано на рис. 3.

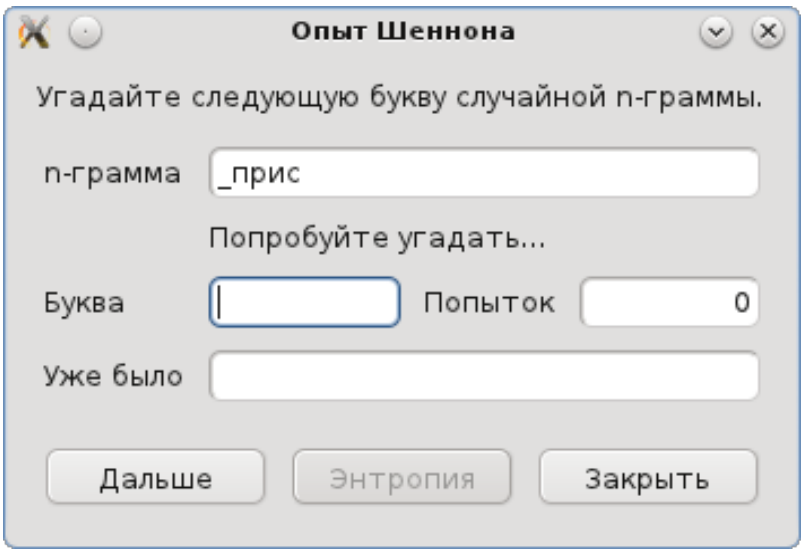



Рис. 3. Опыт Шеннона

Нажатие кнопки «Дальше» до правильного угадывания сбрасывает текущий результат, но не увеличивает число опытов. Нажатие кнопки «Закрыть» сбрасывает все результаты, эксперимент начинается сначала.

Пример распределения и результат расчёта энтропии, полученные в результате многократного повторения опыта, показаны на рис. 4.

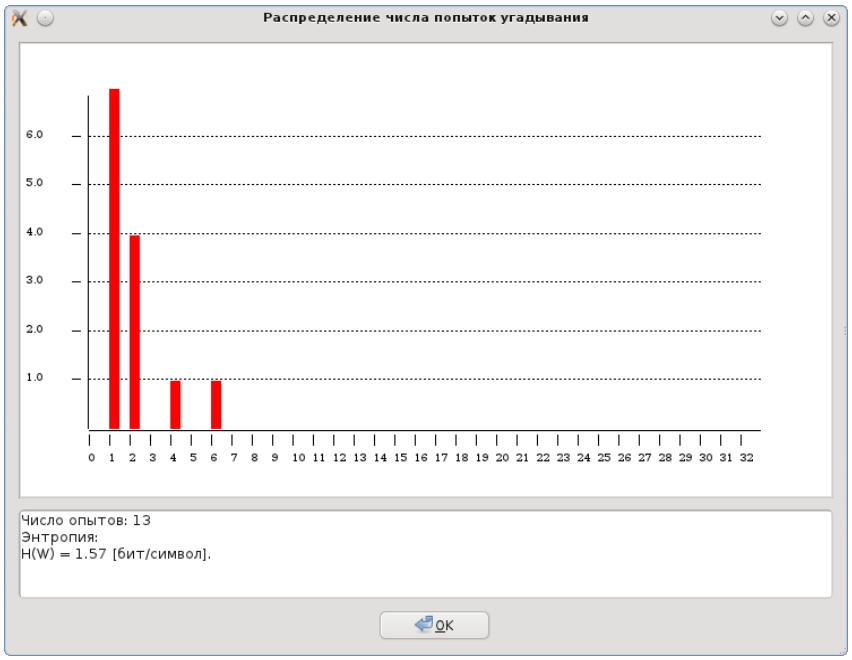



Рис. 4. Определение энтропии в опыте Шеннона

# 3 Литература

1. Шеннон К. Работы по теории информации и кибернетике. — М,: Издательство иностранной литературы, 1963.

2. Яглом И. М., Яглом А. М. Вероятность и информация. — М.: Наука, 1973.